Poster II-50

Middleware Support for Storage and Manipulation of Large Image Datasets in a Grid Environment

Kurc, Tahsin*, Hastings, Shannon, Pan, Tony, Langella, Steve, Catalyurek, Umit, Saltz, Joel
Biomedical Informatics Department, Ohio State University, Columbus, OH, USA

Research and clinical studies in medical imaging are increasingly making use of very large datasets. Examples include time dependent, multi-dimensional, heterogeneous collections of data from multiple imaging sessions and digitized microscopy data obtained by digitizing entire pathology slides or three dimensional datasets acquired from tissue samples through various technologies. In addition, multi-institutional collaboration through sharing of data and resources is increasingly becoming a crucial element in many areas of research ranging from basic science to medicine to engineering. The hardware and networking resources provided by the Grid are ideally suited to support the data sharing, data manipulation and computational requirements associated with advanced imaging applications.

We are developing a services-oriented middleware framework and computational grid services that support efficient use of distributed sets of storage and computation clusters to efficiently store, retrieve, and process imagery in a wide-area environment. The services of the middleware support 1) efficient distribution of image data across potentially heterogeneous collections of storage systems, 2) indexing of large collections of image datasets, each of which may contain very large images, 3) subsetting of large images through spatial and value-based queries, 4) caching of data and analysis results to improve performance of queries in multi client environments, and 5) distributed execution of image processing and analysis operations through collective use of storage and computational clusters in a Grid environment. These services draw from three frameworks we have developed: *Active Data Repository* leverages commonality in processing requirements of data intensive applications to integrate data retrieval and processing on a distributed-memory parallel machine with a disk farm. Its services can be customized to implement user-defined indexing and data processing for applications. *DataCutter* is a component-based framework designed to support coarse grain data flow applications. Through combined task and data parallelism, it enables efficient execution of networks of processing components in a distributed environment. DataCutter has been integrated with several other Grid toolkits such as Globus, Storage Resource Broker, and Network Weather Services for authentication, remote file access, and resource monitoring. *IP4G* is a middleware implemented on DataCutter to support distributed execution of image processing operations. We will describe the overall architecture of the middleware system and will present the application of this system in two projects: 1) Distributed execution of image analysis algorithms for Dynamic Contrast Enhanced MRI studies that involve large number of images. In this project, the distributed execution service of the middleware is employed for executing data segmentation and visualization functions from Insight Segmentation and Registration Toolkit and Visualization Toolkit as well as for performing parameter studies. 2) Remote querying of 3D digitized microscopy images. In this joint project with the NPACI Telescience group, the middleware system is employed to retrieve subsets of large 3D digitized microscopy images from remote storage systems and process them on clusters of PCs.

Federal Grant Support: NIH NIBIB BISTI P20EB000591, Ohio Board of Regents BRTTC BRTT02-0003, NSF EIA-0121161, EIA-0121177, ACI-9619020, ACI-0130437, ACI-0203846, and ACI-9982087, and LLNL B517095 and B500288.